I/O Performance Reproducibility using IO500 Benchmark

Radita Liem
Chair for High Performance Computing, IT Center, RWTH Aachen University

High Performance Computing | IT Center | RWTH AACHEN UNIVERSITY

Motivation for Reproducibility Topic

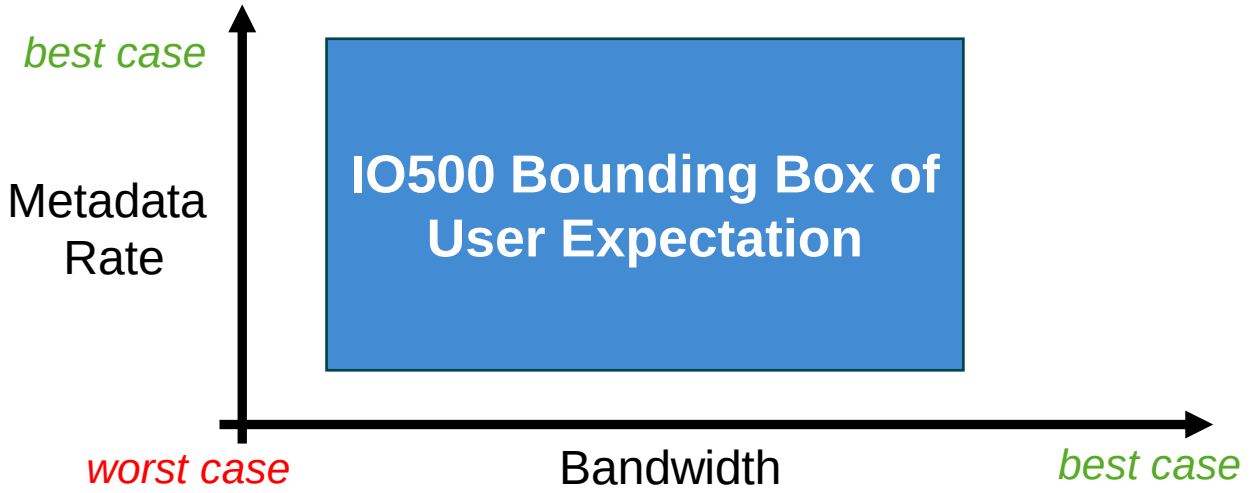
Reproducibility is a major principle in scientific method

Reproducing a paper can be used to onboard and train students

Reproducing IO500 Bounding Box Paper

Paper: **User-Centric System Fault Identification Using IO500 Benchmark** (2021)

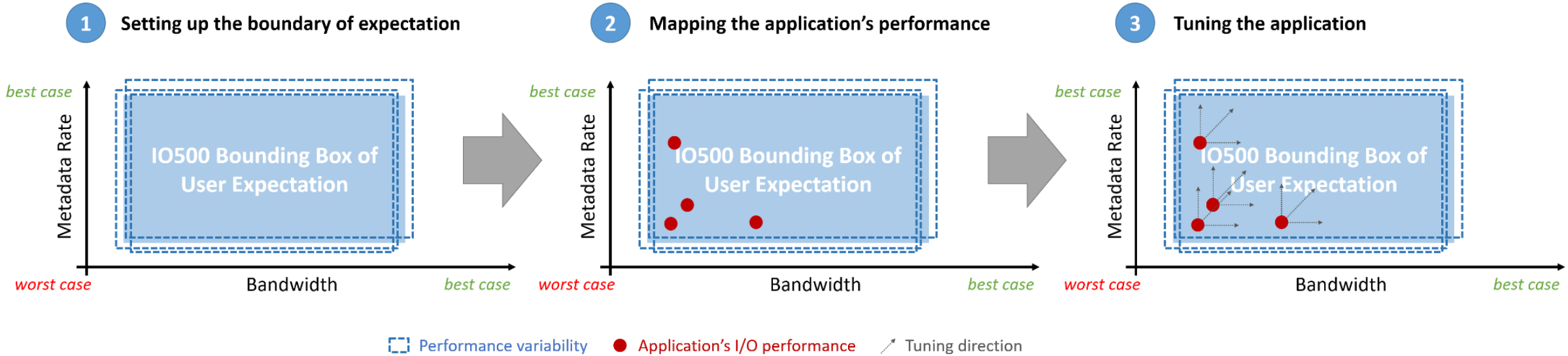
General idea: **IO500 benchmark's mdtest and IOR scenario can be used to form a bounding box of user expectations [4] as illustrated by the figure below**



Worst case scenario is from IOR and mdtest **'hard'** scenario
Best case scenario is from IOR and mdtest '**easy'** scenario

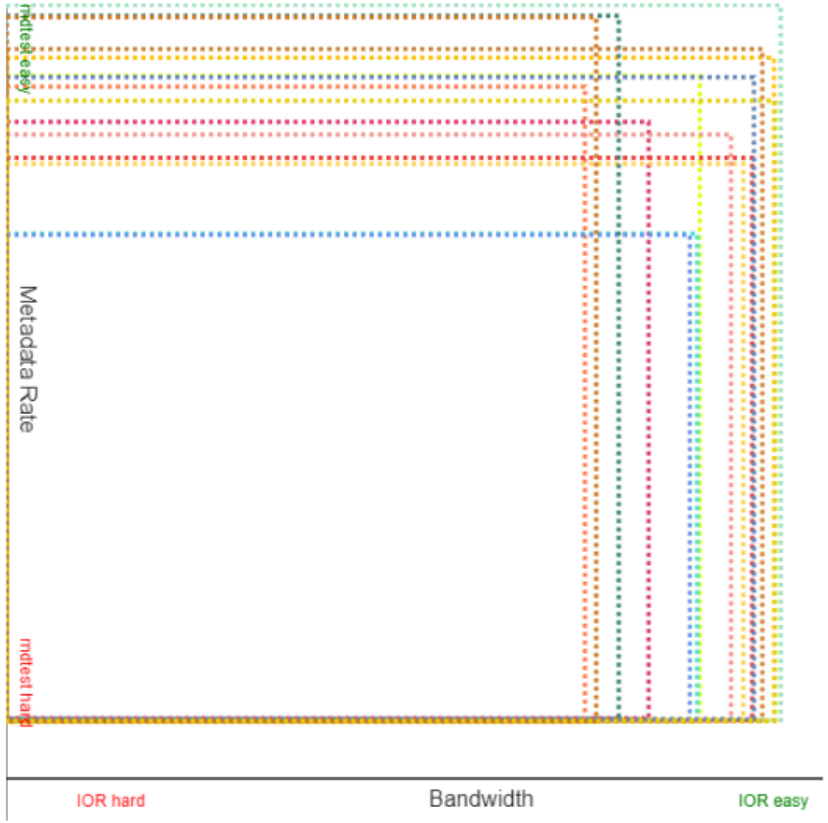
[4] A. Dilger, "IO500 | A storage Benchmark for HPC", 2019. [Online]. Available:https://wiki.lustre.org/images/9/92/LUG2019-IO500_Storage_Benchmark_for_HPC-Dilger.pdf. [Accessed: 02-Mar-2021]

Bounding Box of User Expectation Workflow



Forming Bounding Box of User Expectation

Bounding box of **POSIX** API, each square represents individual run from the same IO configuration



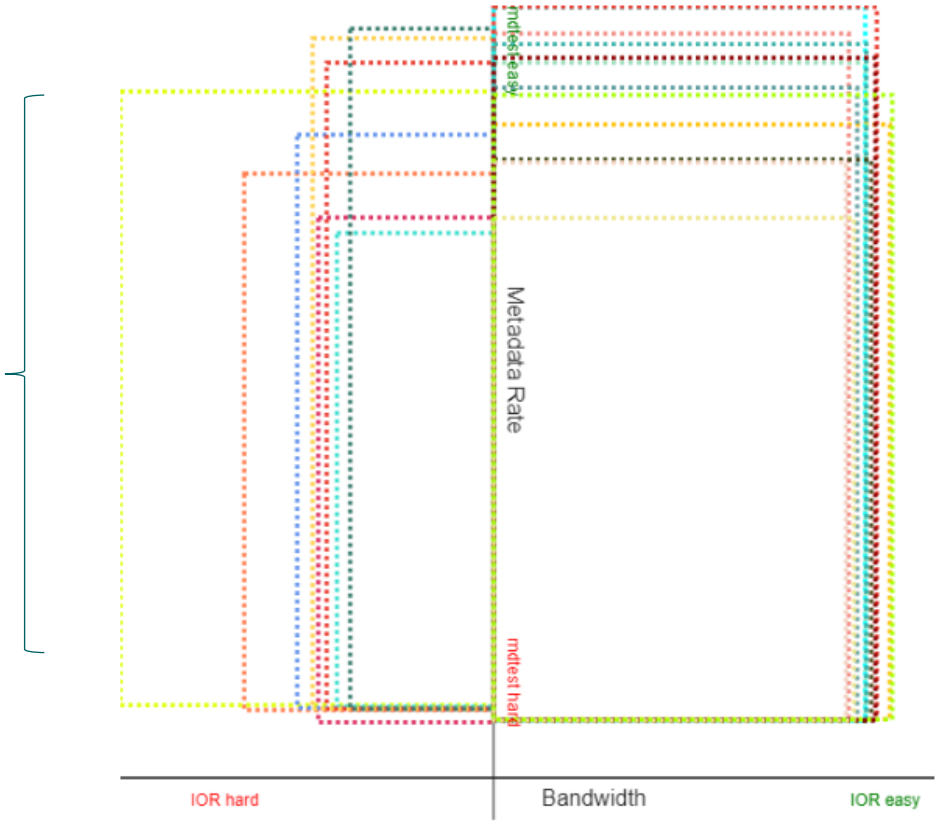
	IOR hard (GiB/s)	IOR easy (GiB/s)	mdtest hard (KIOPS)	mdtest easy (KIOPS)
	0.85	1.88	10.97	145.17
	1.03	1.89	11.26	123.26
	1.10	1.87	10.59	129.86
	0.87	1.90	10.76	135.27
	0.97	1.90	10.64	131.93
	0.94	1.86	11.06	102.35
	0.95	1.86	10.84	102.06
	0.90	1.89	10.98	116.54
	0.90	1.89	11.16	115.40
	1.08	1.90	11.14	143.09
	0.91	1.88	10.80	120.86
	0.90	1.90	11.05	131.65
	0.87	1.88	10.79	136.91

This project is currently displayed in:https://bit.ly/3BhhAFZ

Anomalous Bounding Box

Sometimes, IOR ‘Easy’ score gets lower number than IOR ‘hard’. Broken nodes are the suspect

Bounding box skewed
to the direction of IOR ‘hard’

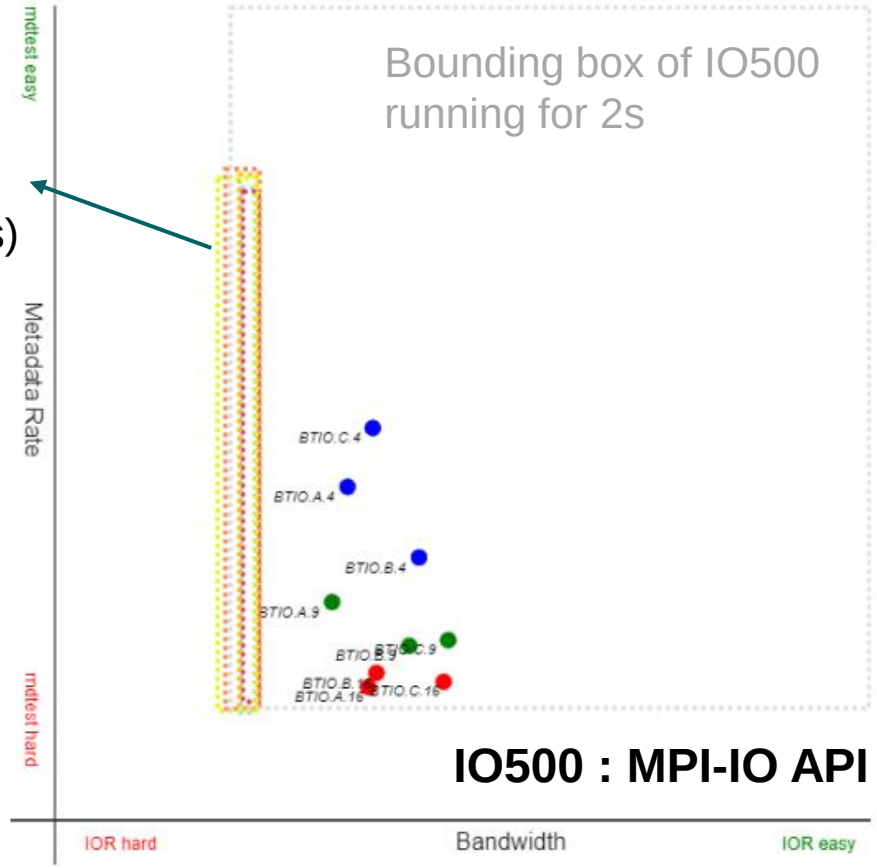


	IOR hard (GiB/s)	IOR easy (GiB/s)	mdtest hard (KIOPS)	mdtest easy (KIOPS)
	0.90	1.89	10.74	135.79
	0.94	0.47	10.50	106.45
	1.14	0.47	12.86	114.73
	0.83	1.89	11.12	124.06
	1.46	0.47	13.78	130.29
	0.88	0.47	13.50	103.47
	0.99	0.47	13.24	122.10
	0.91	0.47	13.04	135.73
	0.95	0.46	13.10	140.41
	0.85	0.47	13.02	142.20
	0.96	1.90	11.00	141.28
	0.86	1.85	10.81	146.24
	0.91	1.87	11.03	131.02

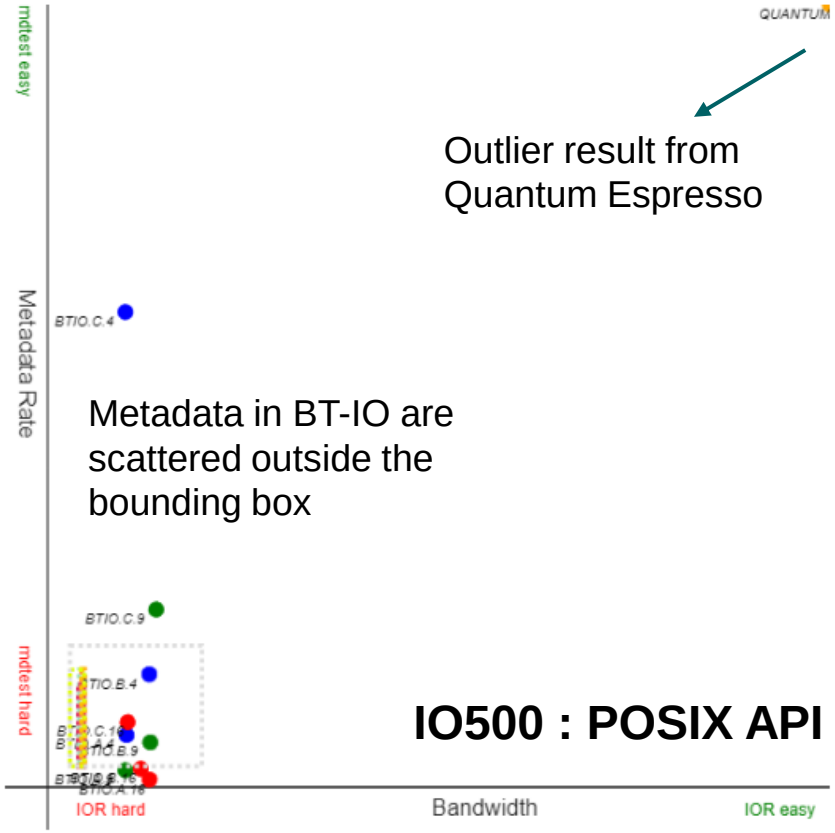
This project is currently displayed in:https://bit.ly/3BhhAFZ

Mapping I/O Performance with Darshan

Bounding boxes of IO500 running with default setup (300s)



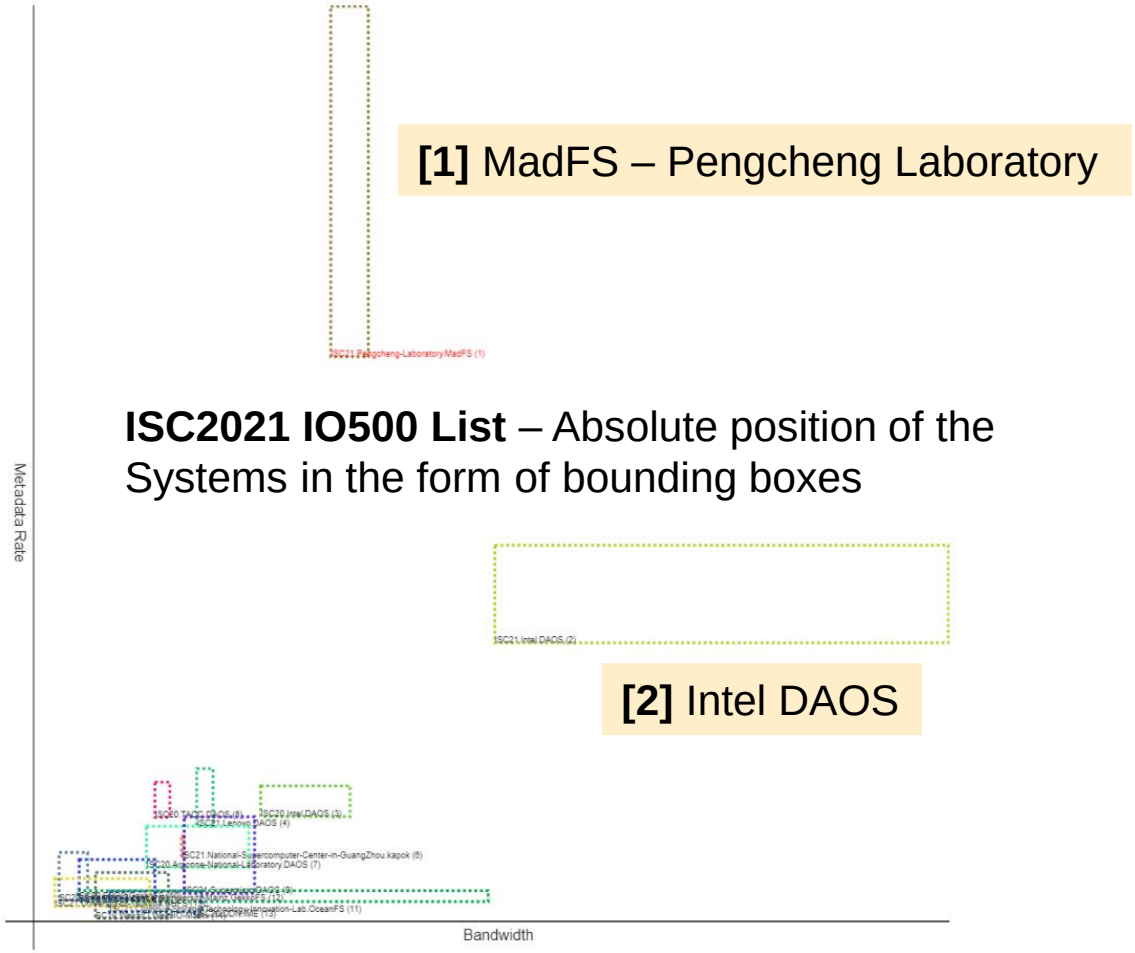
IO500 : MPI-IO API

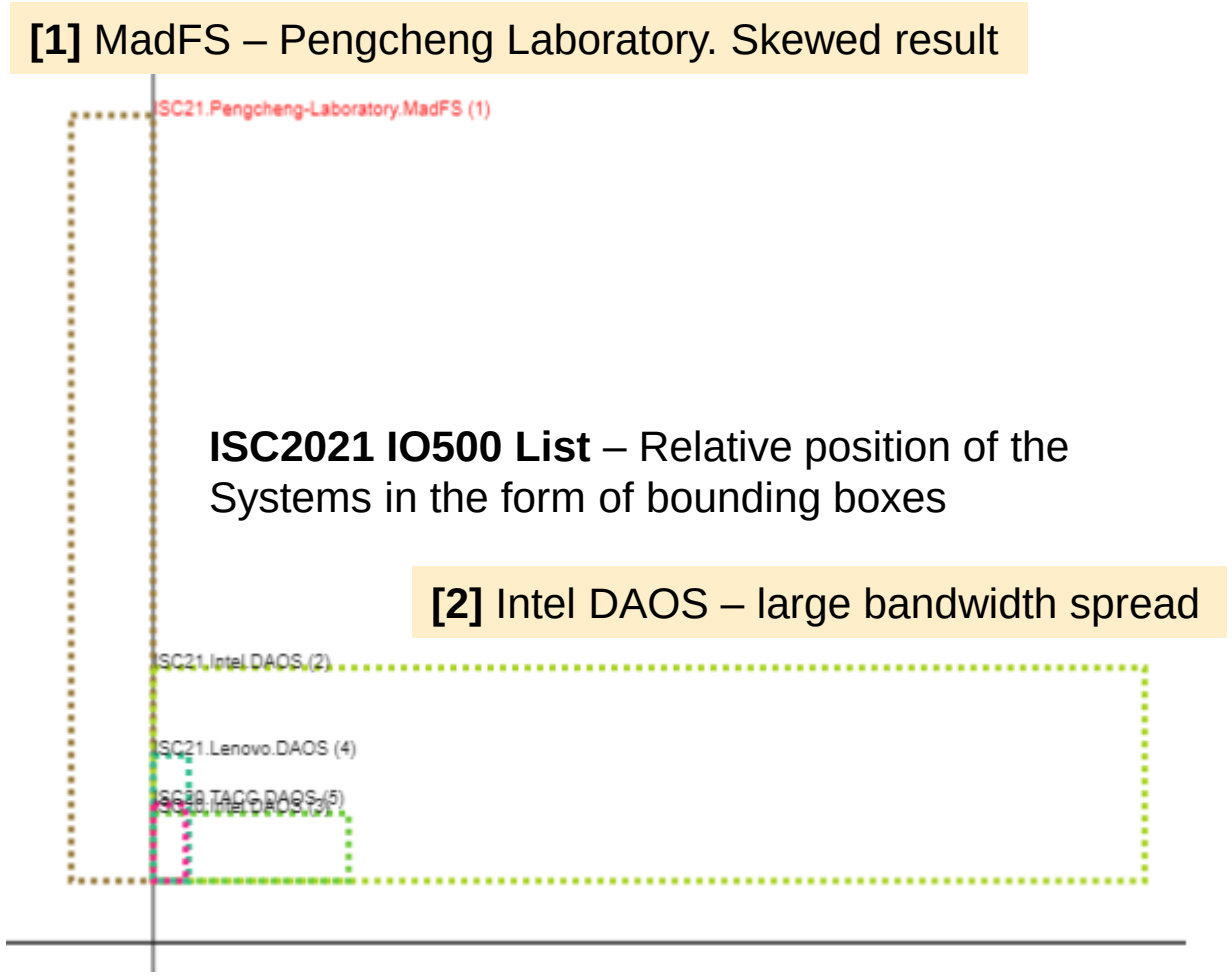


IO500 : POSIX API

This project is currently displayed in:https://bit.ly/3BhhAFZ

Exploring the IO500 List in 2021





Reproducing 2021 Work in 2024

Same system: CLAIX-18 – 4 nodes BeeOND with the same config but different OS

	ISC 2020 benchmark		ISC 2023 benchmark
	CLAIX 2018 (CentOS)	CLAIX 2018 (Rocky Linux)	CLAIX 2018 (Rocky Linux)
find	1468.114386	560.52	361.4813317
ior-easy-read	2.019125	2.14	2.1384143
ior-easy-write	1.731133778	1.761	1.7465284
IOR-EASY	**1.869592332**	**1.941272778**	**1.932563403**
ior-hard-read	1.409629778	0.341	0.3387456
ior-hard-write	0.544298222	0.805	0.7664143
IOR-HARD	**0.875933206**	**0.523932248**	**0.509528676**
mdtest-easy-delete	99.55603122	91.287	86.971433
mdtest-easy-stat	332.9714572	389.352	312.3556426
mdtest-easy-write	66.35817467	102.426	98.1845506
MDTEST-EASY	**130.0537875**	**153.8345383**	**138.682942**
mdtest-hard-delete	9.040069222	8.904	8.8042001
mdtest-hard-read	22.66211533	21.73	20.9965264
mdtest-hard-stat	26.73728556	91.397	89.1813305
mdtest-hard-write	2.601157333	8.068	7.9033316
MDTEST-HARD	**10.92544247**	**19.43505066**	**18.998984**

Slightly better

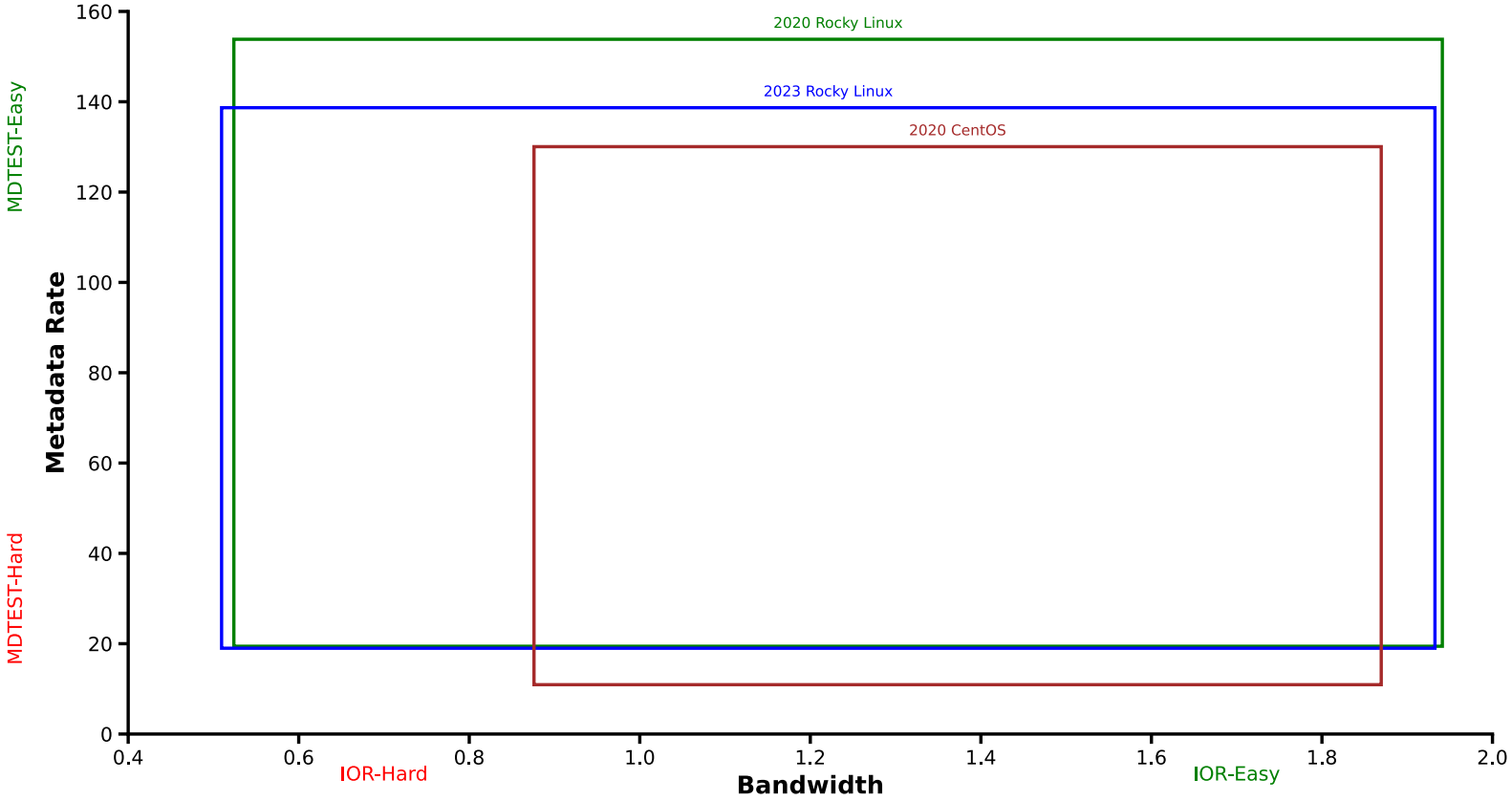
Getting worse

Getting better because of the stat and write getting better

Changes in the Bounding Box

Impact of the OS change or just hardware degradation?

CLAIX-18 cluster at the RWTH Aachen University with 4 nodes of Intel Skylake with 48 cores / node, 192 GB memory and 480 GB SSD / node.

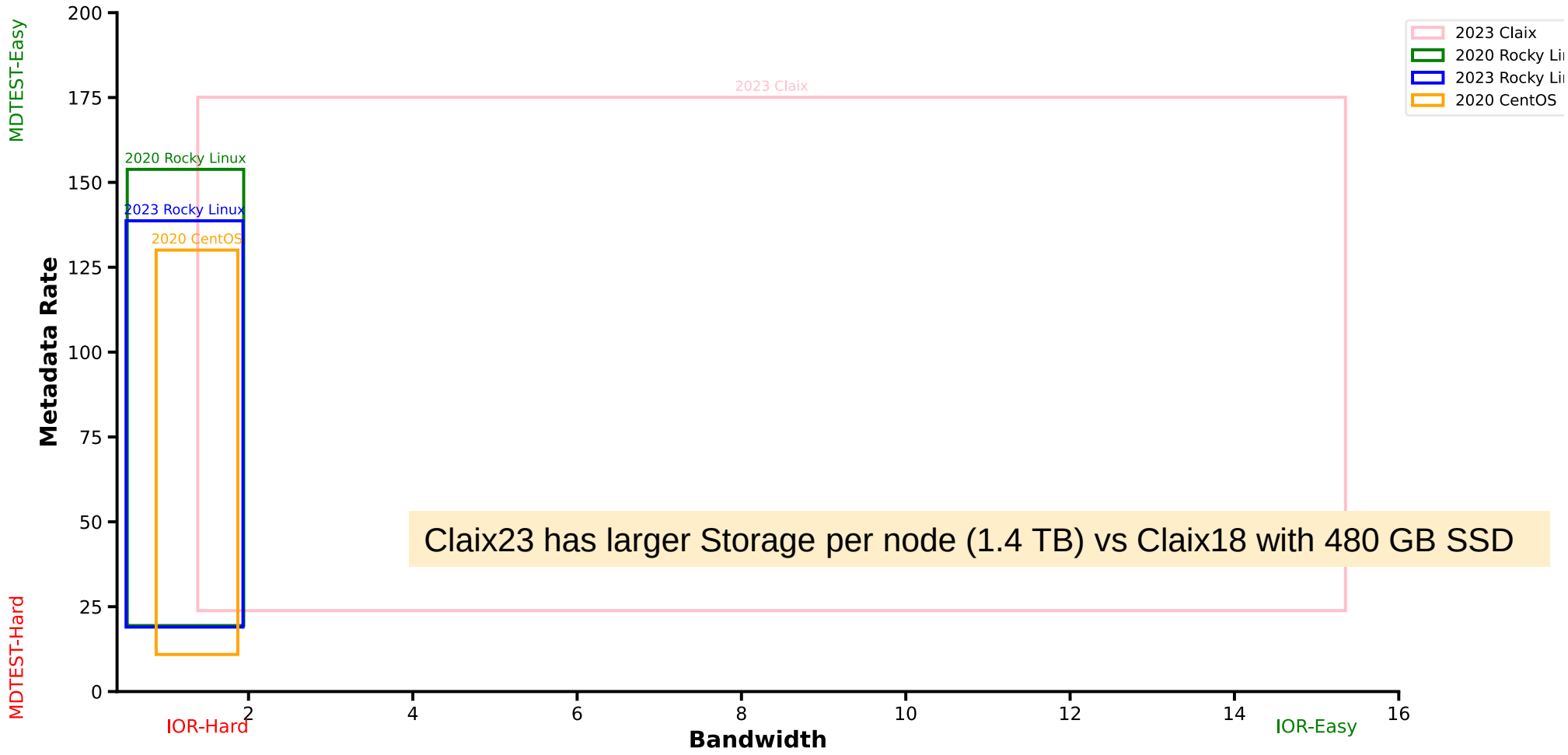


Rerunning IO500 in CLAIX2023

Different system, different result, but by how much?

	ISC 2020 benchmark	ISC 2023 benchmark	
	CLAIX 2018 (Rocky Linux)	CLAIX 2018 (Rocky Linux)	CLAIX 2023 (Rocky Linux)
find	560.52	361.4813317	1018.690577
ior-easy-read	2.14	2.1384143	25.1042156
ior-easy-write	1.761	1.7465284	9.4041188
IOR-EASY	**1.941272778**	**1.932563403**	**15.35235168**
ior-hard-read	0.341	0.3387456	1.001077
ior-hard-write	0.805	0.7664143	1.9128834
IOR-HARD	**0.523932248**	**0.509528676**	**1.381761156**
mdtest-easy-delete	91.287	86.971433	112.9227754
mdtest-easy-stat	389.352	312.3556426	402.9833274
mdtest-easy-write	102.426	98.1845506	117.9598154
MDTEST-EASY	**153.8345383**	**138.682942**	**175.0755941**
mdtest-hard-delete	8.904	8.8042001	11.1692238
mdtest-hard-read	21.73	20.9965264	31.1032606
mdtest-hard-stat	91.397	89.1813305	96.4071642
mdtest-hard-write	8.068	7.9033316	9.6442762
MDTEST-HARD	**19.43505066**	**18.998984**	**23.82826223**

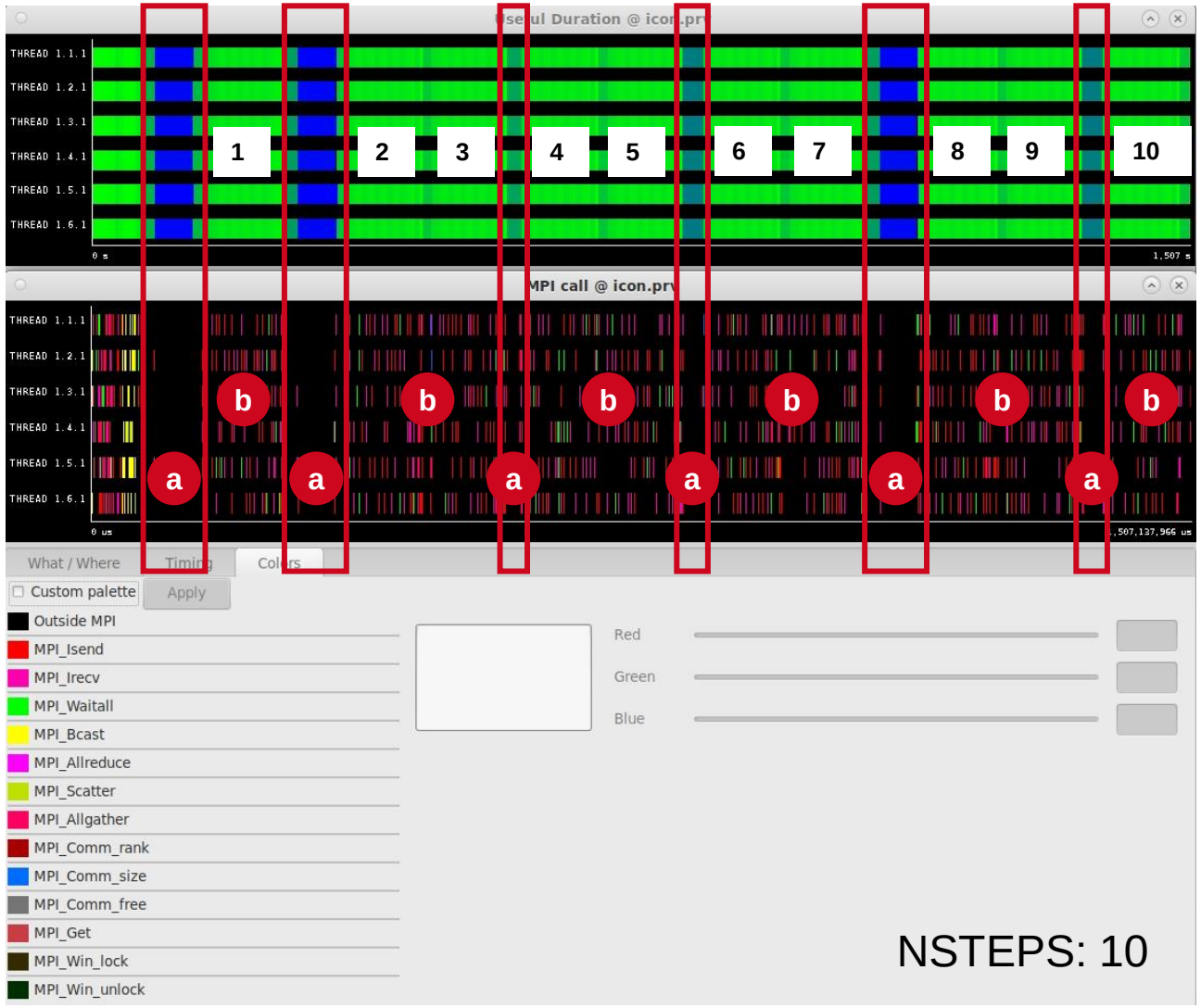
New System (CLAIX-23) vs Old System (CLAIX-18)



Summary & Discussions

- Understanding the changes in performance & interpreting the data
  - There's a decline in the IOR hard performance. Perhaps due to the aging system?
  - New OS creates a performance improvement in all write operations and mdtest-hard-stat. How?
    - The improvement in mdtest-hard-stat might be due to partial dentry caching.
    - Changes in Linux kernel such as changes in the security policy
    - Improvement from the parallel filesystem
- Training for Students:
  - Attracting students who prefer engineering work but in the end, made them read more papers ☺
  - Good documentation is needed
  - I/O literature is quite lacking ☹

ICON Grid R2B7 – General Structure



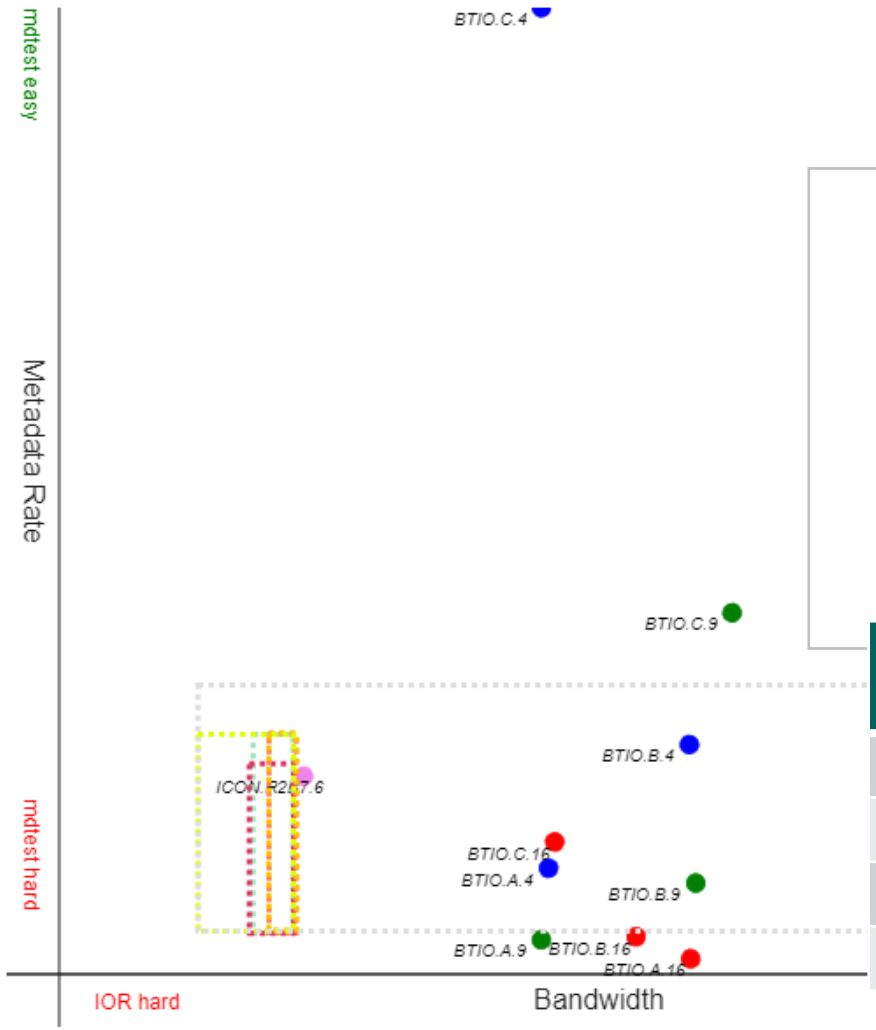
MPI-OpenMP, 6 processes with 8 threads each

Extrae trace with 10 steps iteration.

a. Areas with uninterrupted computation after each step. Closer to blue means longer computation time (~50s) and this areas' duration varies between steps.
b. One step of iteration. This areas filled with MPI asynchronous operations

Dataset provided by Pay Giesselmann (DKRZ)

ICON Grid R2B7 in the CLAIX-18 Bounding Box



	Benchmark	Bandwidth (GiB/s)	Metadata Rate (KIOPS)
	BTIO.A.4	0.99	18.42
	BTIO.A.9	0.97	5.99

ICON's I/O Profile according to Darshan

Filesystem	Data Transferred (MB)	Bandwidth (MB/s)	Runtime
NFS	1913.6	107.06	618.578
BeeGFS	1913.6	504.35	576.454
Lustre	1949.6	122.49	590.859

	GCC + OpenMPI	Intel + IntelMPI	Remarks
Average Runtime	849.75	619.43	27% faster in Intel compilers
Average Energy (PKG)	130145.34	94870.72	27% less energy in Intel
Average Energy (DRAM)	21878.17	16729.11	23.5% less energy in Intel
Average Temperature	53.52	56.00	2.5° hotter in Intel

Thank you!

If you have any question: **Radita Liem** (liem@itc.rwth-aachen.de)